Data Storage and Data Compression

A. M. Odlyzko
Communications Systems Research Section

In this article a sharp upper bound is computed on the best possible data rate achievable as a function of data storage capability in certain very general situations. The result shows that a dramatic increase in rate can be caused by a small increase in storage capability.

I. Introduction

Consider an experiment in which samples x are taken from a sample space X, but no *a priori* probability distribution can be assumed on X. This is the situation, for example, when an experiment is performed for the first time on a distant planet. Further assume that the experimenter, who is separated from the experiment by a communications channel, does not need each value of x exactly, but rather is satisfied with knowing x "approximately." The object of this paper is to discover the relationship between data transmission rate and data storage capacity under these circumstances. We shall discover that a small increase in storage capacity can cause a dramatic increase in data rate.

We formalize the situation as follows. Let X be a set, and let S be a collection of subsets of X. When a sample $x \in X$ is obtained, we assume that the experimenter is satisfied to know only some $A \in S$ such that $x \in A$. Thus if n subsets from S, but no fewer, cover X, $\log_2 n$ bits are required to transmit the outcome of the experiment over a noiseless channel, except for roundoff in the logarithm. If, however, N outcomes $x_1, x_2, \cdots, x_N$ are stored before transmission, and n_N sets of the form $A_1 \times \cdots \times A_N$ cover $X \times \cdots \times X$, then similarly $\log_2 n_N$ bits will suffice to communicate the results of the N outcomes simultaneously; this corresponds to block source encoding of the data. Hence $(1/N) \log_2 n_N$ will be the number of bits per sample required when N samples can be stored. Thus if $n_N < n_1^N$ for some N, the data transmission rate can be improved, preserving the fidelity of each sample. We shall show that under some circumstances n_N actually grows only linearly in N, which implies that a spectacular gain in rate can sometimes be achieved with a small increase in storage. We shall in fact treat the case of simultaneously transmitting the outcomes of several different experiments, since it is no more difficult to handle this more general situation.

II. Results

Suppose that N is a positive integer and that $S_1, \cdots, S_N$ are collections of subsets of $X_1, \cdots, X_N$, respectively, such that n_i subsets belonging to S_i, and no fewer, cover X_i for $i = 1, \cdots, N$. The main result of this paper is that to cover $X_1 \times \cdots \times X_N$ requires no fewer than

$$\sum_{i=1}^{N} (n_i - 1) + 1$$

and no more than

$$\prod_{i=1}^{N} n_i$$

subsets of the form $A_1 \times \cdots \times A_N$, when $A_i \in S_i$ for all i. Moreover, we can choose the X_i and the S_i in such a way that

$$\sum_{i=1}^{N} (n_i - 1) + 1$$

of the specified subsets will cover $X_1 \times \cdots \times X_N$; the X_i, S_i, n_i, can be taken all equal here. Bounds for the spanning number (also called the coefficient of external stability) of a product of graphs are also obtained.

Let S be a collection of subsets of a set X such that $\cup S = X$. Define $c(X; S)$, the covering number of X with respect to S, to be the minimal number of elements of S whose union is X, if this number exists, and infinity if no finite subcollection of S covers X. If $S_1, \cdots, S_N$ are collections of subsets of $X_1, \cdots, X_N$, respectively, define a collection $S_1 \times \cdots \times S_N$ of subsets of the Cartesian product $X_1 \times \cdots \times X_N$ by

$$S_1 \times \cdots \times S_N = \{A_1 \times \cdots \times A_N : A_i \in S_i, i = 1, \cdots, N\}$$

This paper is concerned with the dependence of

$$c(X_1 \times \cdots \times X_N; S_1 \times \cdots \times S_N)$$

on the $c(X_i; S_i)$. We will restrict ourselves to the case in which all the $c(X_i; S_i)$ are finite, since otherwise $c(X_1 \times \cdots \times X_N; S_1 \times \cdots \times S_N) = \infty$.

If the sets A_{i,j_i} for $j_i = 1, \cdots, c(X_i; S_i)$ cover X_i for $i = 1, \cdots, N$, then the sets $A_{i,j_1} \times \cdots \times A_{N,j_N}$ cover $X_1 \times \cdots \times X_N$. We then easily obtain the upper bound

$$c(X_1 \times \cdots \times X_N; S_1 \times \cdots \times S_N) \leq \prod_{i=1}^{N} c(X_i; S_i)$$

Moreover, this bound cannot be improved, since equality holds whenever an X_i consists of n_i points and the subsets in S_i consist of single points. Now if $n_1, \cdots, n_N$ are positive integers, let

$$L(n_1, \cdots, n_N) = \min \{c(X_1 \times \cdots \times X_N; S_1 \times \cdots \times S_N) : c(X_i; S_i) = n_i, 0 = 1, \cdots, N\}$$

We will prove (Theorem 1) the rather surprising result that

$$L(n_1, \cdots, n_N) = \sum_{i=1}^{N} (n_i - 1) + 1$$

The last part of this paper deals with direct products of graphs. Let us recall that if $G_1, \cdots, G_N$ are graphs, their product $G_1 \times \cdots \times G_N$ is defined as the graph having as its nodes ordered N-tuples $(a_1, \cdots, a_N)$, when a_i is a node of G_i for $i = 1, \cdots, N$, and whose two nodes $(a_1, \cdots, a_N)$ and $(b_1, \cdots, b_N)$ are connected by an edge if and only if they are distinct and for each $i = 1, \cdots, N$ either $a_i = b_i$ or a_i is connected to b_i by an edge of G_i. A *talon* is defined as a node together with all nodes that are connected to it by an edge. The spanning number $\beta(G)$ (also called the coefficient of external stability (Ref. 1) of a graph G is the smallest number of talons that cover (contain every node of) G. It is easy to see that if $G_1, \cdots, G_N$ are graphs,

$$\beta(G_1 \times \cdots \times G_N) \leq \prod_{i=1}^{N} \beta(G_i)$$

We will prove that

$$\beta(G_1 \times \cdots \times G_N) \geq \sum_{i=1}^{N} (\beta(G_i) - 1) + 1$$

and that given any positive integers $n_1, \cdots, n_N$ there are graphs $G_1, \cdots, G_N$ with $\beta(G_i) = n_i$ for $i = 1, \cdots, N$ such that

$$\beta(G_1 \times \cdots \times G_N) = \sum_{i=1}^{N} (n_i - 1) + 1$$

(Again if $n_i = n$, all i, then the G can be taken as equal.)

This paper may be regarded as the dual to the work of Erdős, McEliece, and Taylor (Ref. 2). If $\alpha(G)$ is the independence numbers (maximal numbers of nodes such that no two are connected by an edge, also called the coefficient of internal stability [Ref. 1]) of a graph G, then they proved that

$$\alpha(G_1 \times \cdots \times G_N) \leq M(\alpha(G_1) + 1, \cdots, \alpha(G_N) + 1) - 1$$

where $M(k_1, \cdots, k_N)$ is the Ramsey number of $k_1, \cdots, k_N$. It is easy to see that

$$\alpha(G_1 \times \cdots \times G_N) \geq \prod_{i=1}^{N} \alpha(G_i)$$

Moreover, they showed that given positive integers $n_1, \cdots, n_N$, it is possible to find graphs $G_1, \cdots, G_N$ such that $\alpha(G_i) = n_i$ for all i and

$$\alpha(G_1 \times \cdots \times G_N) = M(n_1 + 1, \cdots, n_N + 1) - 1$$

and the equality remark again holds.

These theorems correspond to our results on the spanning number of a product of graphs. In addition, however, they can be interpreted as results on packing. If S is a collection of subsets of a set X, define $p(X; S)$ to be the largest number of elements of S such that any two have an empty intersection (we will assume that this number is finite). We can associate with X and S a graph G, where the nodes of G will be the elements of S and two nodes will be joined by an edge if and only if this intersection (as subsets of X) is nonempty. Moreover, given any graph G we can find a set X and a collection S of subsets of X such that the graph associated with X and S will be isomorphic to G. The importance of this correspondence is that if G is the graph associated with X and S then

$$\alpha(G) = p(X; S)$$

and the theorem of Erdős, McEliece, and Taylor says that

$$p(X_1 \times \cdots \times X_N; S_1 \times \cdots \times S_N) \leqq M(p(X_1, S_1) + 1, \cdots, p(X_N, S_N) + 1) - 1$$

and that this bound is the best one possible.

Lemma 1. *If $n_1, \cdots, n_N$ are positive integers, then*

$$L(n_1 + 1, n_2, \cdots, n_N) \geqq L(n_1, n_2, \cdots, n_N) + 1$$

Proof. Suppose that $S_1, \cdots, S_N$ are collections of subsets of $X_1, \cdots, X_N$, respectively, such that $c(X_1; S_1) = n_1 + 1$, $c(X_i; S_i) = n_i$ for $i = 2, \cdots, N$, and $c(X_1 \times \cdots \times X_N; S_1 \times \cdots \times S_N) = L(n_1 + 1, n_2, \cdots, n_N)$. Let $A_{1,j} \times \cdots \times A_{N,j}$ for $j = 1, \cdots, L(n_1 + 1, n_2, \cdots, n_N)$ be a minimal covering of $X_1 \times \cdots \times X_N$ by subsets from $S_1 \times \cdots \times S_N$. Consider

$$X_1' = X_1 - A_{1,1} = \{x \in X_1 : x \notin A_{1,1}\}$$

and

$$S_1' = \{A - A_{1,1} : A \in S_1\}$$

The sets $(A_{1,j} - A_{1,1}) \times A_{2,j} \times \cdots \times A_{N,j}$ for

$$j = 2, \cdots, L(n_1 + 1, n_2, \cdots, n_N)$$

belong to $S_1' \times S_2 \times \cdots \times S_N$ and cover $X_1' \times X_2 \times \cdots \times X_N$, so that

$$L(n_1 + 1, n_2, \cdots, n_N) - 1 \geqq c(X_1' \times X_2 \times \cdots \times X_N; S_1' \times S_2 \times \cdots \times S_N) \tag{1}$$

But $c(X_1'; S_1') = n_1$ or $n_1 + 1$, so that

$$c(X_1' \times X_2 \times \cdots \times X_N; S_1' \times S_2 \times \cdots \times S_N) \geqq L(n_1, n_2, \cdots, n_N) \tag{2}$$

or

$$c(X_1' \times X_2 \times \cdots \times X_N; S_1' \times S_2 \times \cdots \times S_N) \geqq L(n_1 + 1, n_2, \cdots, n_N) \tag{3}$$

Combining Eqs. (1) and (3) leads to an immediate contradiction, and hence Eq. (2) must hold. The lemma now follows from Eqs. (1) and (2).

Theorem 1. *If $n_1, \cdots, n_N$ are positive integers, then*

$$L(n_1, \cdots, n_N) = \sum_{i=1}^{N} (n_i - 1) + 1 \tag{4}$$

Proof. Lemma 1, together with the fact that $L(k_1, \cdots, k_N)$ is a symmetric function and that $L(1, \cdots, 1) = 1$, implies that

$$L(n_1, \cdots, n_N) \geqq \sum_{i=1}^{N} (n_i - 1) + 1 \tag{5}$$

To prove Eq. (4) it will therefore suffice if we exhibit X_i and S_i such that $c(X_i; S_i) = n_i$ for $i = 1, \cdots, N$ and

$$c(X_1 \times \cdots \times X_N; S_1 \times \cdots \times S_N) \leqq \sum_{i=1}^{N} (n_i - 1) + 1$$

Let

$$n = \sum_{i=1}^{N} (n_i - 1) + 1$$

and define, for each $i = 1, \cdots, N$

$$X_i = \left\{1, 2, \cdots, \binom{n}{n - n_i + 1}\right\}$$

where $\binom{a}{b}$ are the binomial coefficients. The family S_i will consist of n subsets $A_{i,1}, \cdots, A_{i,n}$ formed as follows: number the

$$\binom{n}{n - n_i + 1}$$

possible $(n - n_i + 1)$ subsets (subsets with $n - n_i + 1$ elements) of S_i from 1 to

$$\binom{n}{n - n_i + 1}$$

and assign the integer i to each of the $n - n_i + 1$ sets in the ith collection. This way each element of X_i will belong to exactly $n - n_i + 1$ of the sets $A_{i,1}, \cdots, A_{i,n}$ and each $n - n_i + 1$ of these sets will have an element in common.

Let any $n_i - 1$ sets from S_i be given. By their definition the remaining $n - n_i + 1$ sets have a point $x \in X_i$ in common. Since x belongs to exactly $n - n_i + 1$ sets from S_i, it does not belong to any of the given $n_i - 1$ sets. Therefore these $n_i - 1$ sets do not cover X_i. On the other hand, if any n_i subsets from S_i are given, then every point $x \in X_i$ is contained in at least one of them since it is contained in $n - n_i + 1$ subsets from S_i and aside from the given ones there are only $n - n_i$ subsets remaining. Therefore $c(X_i; S_i) = n_i$.

Now consider the sets

$$\widetilde{A}_j = A_{1,j} \times \cdots \times A_{N,j} \qquad \text{for } j = 1, \cdots, n$$

Suppose $x = (x_1, \cdots, x_N) \in X_1 \times \cdots \times X_N$. We know that if i is fixed, then $x_i \notin A_{i,j}$ holds for exactly $(n_i - 1)$ values of j. Since $x \notin \widetilde{A}_j$ occurs only when $x_i \notin A_{i,j}$ for at least one i_1, it cannot occur for more than

$$\sum_{i=1}^{N} (n_i - 1)$$

values of j. Since there are

$$n = \sum_{i=1}^{N} (n_i - 1) + 1$$

sets $\widetilde{A}_j$, we conclude that $x \in \widetilde{A}_j$ for at least one j. Thus the sets $\widetilde{A}_j$ provide a cover for $X_1 \times \cdots \times X_N$, and hence

$$c(X_1 \times \cdots \times X_N; S_1 \times \cdots \times S_N) \leq n$$

This completes the proof of the theorem.

An interesting fact emerges from an examination of the above theorem. If $n_1 = n_2 = \cdots = n_N$ then a minimal covering with $L(n_1, \cdots, n_N)$ sets is obtained by taking all the X_i and likewise all the S_i equal. To prove this, just check that the above construction provides the desired covering, *mutatis mutandis.*

Theorem 2. *If $G_1, \cdots, G_N$ are graphs and $\beta(G)$ denotes the spanning number of a graph G, then*

$$\beta(G_1 \times \cdots \times G_N) \geq \sum_{i=1}^{N} (\beta(G_i) - 1) + 1 \qquad (6)$$

Moreover, given any positive integers $n_1, \cdots, n_N$ there are graphs $G_1, \cdots, G_N$ with $\beta(G_i) = n_i$ for $i = 1, \cdots, N$ and

$$\beta(G_1 \times \cdots \times G_N) = \sum_{i=1}^{N} (n_i - 1) + 1 \qquad (7)$$

Proof. Inequality (6) follows immediately from Eq. (5), since for each i we can take X_i to consist of the nodes of G_i and S_i to consist of talons of G_i, so that $\beta(G_i) = c(X_i; S_i)$ and $\beta(G_1 \times \cdots \times G_N) = c(X_1 \times \cdots \times X_N; S_1 \times \cdots \times S_N)$. In fact, Eq. (4) follows from Eqs. (5) and (7). Since the proof of Eq. (7) is considerably more complicated than that of Eq. (4), however, we treat products of graphs separately from products of arbitrary sets. Nevertheless, there will be very little duplication in the two proofs, since we will often refer to arguments used in proving Theorem 1. The proof of the last part of this theorem will be based on that of Theorem 1. To carry it out we will use the auxiliary result below (we might note that the bound $n > 2k + 10$ is introduced to simplify the proof and is not the best possible.) The proof of this lemma will be given at the end.

Lemma 2. *If n and k are positive integers such that $n > 2k + 10$, then there exists a graph G with $\beta(G) = k$ which has a subset of n talons such that every node of G belongs to at least $n - k + 1$ of these specified n talons.*

To complete the proof of Theorem 2 let $n_1, \cdots, n_N$ be given positive integers. Let us choose the positive integer r such that if

$$n = \sum_{i=1}^{N} (n_i - 1) + 1 + r$$

then $n > 2n_i + 10$ for $i = 1, \cdots, N$ and $n > 14$. Define $n_j = 2$ for $j = N + 1, \cdots, N + r$. By Lemma 2 there exists for each i, from 1 to $N + r$, a graph G_i with $\beta(G_i) = n_i$ and which has a subset of n talons, $A_{i,1}, \cdots, A_{i,n}$, such that every node of G_i belongs to at least $n - n_i + 1$ of these talons. It now follows, as in Theorem 1, that the n talons

$$\widetilde{A}_j = A_{1,j} \times \cdots \times A_{N+r,j}, \qquad j = 1, \cdots, n$$

cover $G_1 \times \cdots \times G_{N+r}$, which implies that

$$\beta(G_1 \times \cdots \times G_{N+r}) \leq n = \sum_{i=1}^{N} (n_i - 1) + 1 + r \qquad (8)$$

But from Eq. (6) we find that

$$\beta(G_1 \times \cdots \times G_{N+r}) \geqq \beta(G_1 \times \cdots \times G_N) + \beta(G_{N+1} \times \cdots \times G_{N+r}) - 1 \tag{9}$$

Since

$$\beta(G_1 \times \cdots \times G_N) \geqq \sum_{i=1}^{N} (n_i - 1) + 1$$

and

$$\beta(\tilde{G}_{N+1} \times \cdots \times G_{N+r}) \geqq r + 1$$

The inequalities (8) and (9) imply that

$$\beta(G_1 \times \cdots \times G_N) = \sum_{i=1}^{N} (n_i - 1) + 1$$

which completes the proof of Theorem 2.

Proof of Lemma 2. Let n and k satisfy conditions of the lemma. We consider two cases.

Case 1: k odd. Let the nodes of G be $\{1, \cdots, \binom{n}{n-k+1}\}$. As in Theorem 1 we construct subsets $A_1, \cdots, A_n$ such that every $n - k + 1$ of these has a node in common and every node in G is in exactly $n - k + 1$ of these subsets. Let us look at the incidence matrix (a_{ij}) $(i = 1, \cdots, n$ and $j = 1, \cdots, \binom{n}{n-k+1})$ of this configuration, where $a_{ij} = 1$ if $j \in A_i$ and $a_{ij} = 0$ otherwise. Since every possible arrangement of $n - k + 1$ 1's and $k - 1$ 0's occurs exactly once in some column, we can permute the columns of this matrix until the first n of them form a circulant submatrix A of the form

$$A = \left[\begin{array}{l} 1\;\overbrace{0\ldots\ldots0}^{\frac{k-1}{2}}\;1\ldots\ldots1\;\overbrace{0\ldots\ldots0}^{\frac{k-1}{2}} \\ 0\;1\ldots\ldots\ldots0\;1\ldots\ldots1\;0\ldots.0 \\ \\ 0\ldots\ldots0\;1\ldots\ldots1\;0\ldots\ldots0\;1 \end{array}\right] \tag{10}$$

Note that this construction can be carried out because each row (and column) has at least three 1's in it (if they each had two, some two columns would be identical), and because k is odd. For our purposes the most important feature of A is that it is symmetric.

Let us now consider again the full incidence matrix (a_{ij}), but in its permuted form so that A forms the first n columns. For each i from 1 to n connect the node i to the node j (for $i \neq j$) by an edge if and only if $a_{ij} = 1$. This defines our graph G. Since A is symmetric, the talon with center at $i\,(i \leqq n)$ consists precisely of these nodes j for which $n_{ij} = 1$. We will call them the "large talons." In addition we have many "small talons" which have centers at i, when $i > n$. A small talon with center at i contains the node i and $n - k + 1$ nodes j for $j \leqq n$. Since each node of G is in $n - k + 1$ of the large talons, we only need to show $\beta(G) = k$. Since any k large talons cover G (as shown in Theorem 1), we know that $\beta(G) \leqq k$.

Assume $\beta(G) < k$. Since $\beta(G) \geqq 1$, this implies $k \geqq 2$. Now $\beta(G) < k$ says that some $k - 1$ talons cover G. Suppose m of them are large talons. Then $m \leqq k - 2$, since we know from Theorem 1 that it requires k large talons to cover G. Moreover, $k \geqq 3$, since if $k = 2$ then there are no small talons. We can choose $n - k + 1$ large talons from the $n - m$ remaining ones in $\binom{n-m}{n-k+1}$ ways. Since each $n - k + 1$ large talons have a unique node in common, which is not in any other large talon, this gives $\binom{n-m}{n-k+1}$ nodes not covered by the m given large talons.

Suppose $m = k - 2$. This means that one small talon has to cover $\binom{n-k+2}{n-k+1} = n - k + 2$ nodes. Since exactly one node i with $i > n$ belongs to any small talon, the m given large talons had to leave $n - k + 1$ nodes j with $j \leqq n$ not covered. But any single large talon covers $n - k + 1$ of the nodes j with $j \leqq n$, and therefore our $m = k - 2 \geqq 1$ large talons leave uncovered no more than $k - 1$ nodes. Since $n > 2k + 10$, this leads to a contradiction.

Suppose $m < k - 2$. The $k - 1 - m$ small talons cover at most $(k - 1 - m)(n - k + 2)$ nodes, while the m large talons have $\binom{n-m}{n-k+1}$ nodes not covered. It will then suffice to prove that

$$\binom{n-m}{n-k+1} > (k - m - 1)(n - k + 2)$$

Now if $m = k - 3$,

$$\binom{n-m}{n-k+1} = \frac{(n-m)(n-m-1)}{2} > 2(n - m - 1)$$

since $n > 2k + 10 > m + 4$. Let $t = k - m - 3$. Then

$$\binom{n-m}{n-k+1} = \binom{n-m}{t+2} = \frac{(n-m)\cdots(n-k+2)}{(t+2)!}$$

$$\geqq (n-k+2)\frac{(n-k+3)^{t+1}}{(t+2)!}$$

$$> (n-k+2)\frac{(k+13)^{t+1}}{(t+2)!}$$

$$> (n-k+2)(k-m-1)\frac{(k+13)^{t}}{(t+2)!}$$

$$> (n-k+2)(k-m-1)$$

for $1 \leqq t \leqq k-3$, since in that range

$$\frac{(k+13)^t}{(t+2)!} > 1$$

(easy induction proof). This completes the proof when k is odd.

Case 2: k even. Consider the same construction as in case 1, but with n and k replaced by $n + 1$ and $k + 1$, respectively. In the $(n + 1) \times \binom{n+1}{n-k+1}$ incidence matrix (a_{ij}) which has its first $n + 1$ columns in the form (10) delete the first row and the first column. The first n columns of the resulting $n \times \binom{n+1}{n-k+1}$ incidence matrix (b_{ij}) form a symmetric submatrix. The graph G will have as its nodes $\{1, \cdots, \binom{n+1}{n-k+1} - 1\}$. If $i \leqq n$, we connect i to j $(i \neq j)$ by an edge if and only if $b_{ij} = 1$. It is easily seen that every node of G belongs to either $n - k + 1$ or $n - k + 2$ of the n talons with centers at $1, \cdots, n$. The proof of case 1 shows that $\beta(G) = k$, *mutatis mutandis*.

As a concluding remark we would like to pose another interesting problem. If G is a graph, $\gamma(G)$, the clique covering number, is defined to be the minimal number of cliques (complete subgraphs) which cover G. If $G_1, \ldots, G_N$ are graphs, then clearly

$$\gamma(G_1 \times \cdots \times G_N) \leqq \prod_{i=1}^{N} \gamma(G_i)$$

and this bound cannot be improved. The question of obtaining the best possible lower bound for $\gamma(G_1 \times \cdots \times G_N)$ in terms of the $\gamma(G_i)$ is unsolved. The answer is different from that for spanning numbers, since if $\gamma(G_1) = 2$, then $\gamma(G_1 \times G_2) = 2\gamma(G_2)$. Combined with a modified version of Lemma 1 this provides a general lower bound for $\gamma(G_1 \times \cdots \times G_N)$, but it is not known whether it is the best possible. In particular, it is not known whether two sequences $\{G^{(n)}\}$ and $\{H^{(n)}\}$ of graphs can be found such that $\gamma(G^{(n)}) = \gamma(H^{(n)}) = n$ and $\gamma(G^{(n)} \times H^{(n)}) = o(n^2)$.

Acknowledgment

I would like to thank R. J. McEliece and E. C. Posner for proposing these problems and for their help. I would also like to thank H. Taylor for many stimulating conversations and especially for suggesting the basic construction used in proving Theorem 2.

References

1. Berge, C., *The Theory of Graphs and Its Applications*, Wiley & Johns, Inc., New York, 1962.

2. Erdős, P., McEliece, R. J., and Taylor, H., "Ramsey Bounds for Graph Products," *Pacific J. Math*, Vol. 37, pp. 45–46, 1971.